# Integração e Processamento Analítico de Informação Project

## Stage I: Data Analysis


**Group 1**

Tiago Rodrigues nº49593

Ivo Oliveira nº50301

Martim Silva nº51304

Alexandre Sobreira nº59451


2022/2023

# Table of Contents

# 1 - Identify the data sources used

This project will aim to aid a big e-commerce chain in improving its business by analysing various datasets that are related directly or indirectly to it. With this goal in mind, four data sets will be analysed together. The four datasets are Global Superstore Orders 2016; Global Superstore Returns 2016; Superstore Dataset; US Holiday Dates (2004-2021); Highest GDP Counties in the USA. All the analysis in this step was done using python 3 with the aid of packages such as pandas, NumPy, seaborn and matplotlib. The four datasets will be described in the following steps.

## 1.1 - Global Superstore Orders 2016.xlsx

This dataset can be found in the open source "data.world" (https://data.world/tableauhelp/superstore-data-sets). It has a total of 8.08 MB in xlsx format, containing two tables, Orders and People, which we will call Sellers. This dataset includes information regarding the sales of the large-scale e-commerce chain.

### 1.1.1 - Orders

The initial dataset table is called "Orders," and it serves as the primary source of information for this project. The table includes 51290 rows and 24 columns, which are outlined below.

*Table 1 - Orders table column number (#), name, description and data type*

| # | Column | Description | Data type |
|---|---|---|---|
| **1** | **Row ID** | Unique ID for each row. | Int64 |
| **2** | **Order ID** | Unique Order ID for each Customer. | Object |
| **3** | **Order Date** | Order Date of the product. | Datetime64[ns] |
| **4** | **Ship Date** | Shipping Date of the Product. | Datetime64[ns] |
| **5** | **Ship Mode** | Shipping Mode specified by the customer. | Object |
| **6** | **Customer ID** | Unique ID to identify each customer. | Object |
| **7** | **Customer Name** | Name of the Customer. | Object |
| **8** | **Segment** | The segment where the customer belongs. | Object |
| **9** | **Postal Code** | Postal Code of every Customer | Float64 |
| **10** | **City** | City of residence of the seller. | Object |
| **11** | **State** | State of residence of the seller. | Object |
| **12** | **Country** | Country of residence of the seller. | Object |
| **13** | **Region** | The region where the seller belongs. | Object |
| **14** | **Market** | Global location by Market | Object |
| **15** | **Product ID** | Unique ID of the Product | Object |
| **16** | **Category** | The Category of the product ordered. | Object |
| **17** | **Product Name** | Name of the Product | Object |
| **18** | **Sales** | Sales of the Product. | Float64 |
| **19** | **Quantity** | Quantity of the Product. | Int64 |
| **20** | **Discount** | Discount provided. | Float64 |
| **21** | **Profit** | Profit/Loss incurred. | Float64 |
| **23** | **Shipping Cost** | Cost per shipment | Float64 |
| **24** | **Order Priority** | Level of priority for the order | Object |

Table 1 shows that the Orders table comprises various data types, including order and shipment dates, customer and seller information (e.g., region), and product details such as category and profit.

For clarity, the variables in the Orders table were divided into categorical and numerical categories to aid in their description, which can be seen below.

*Table 2 - Orders table categorical variables description with a column name, number of unique instances and example of content.*

| Column | Unique | Example |
|---|---|---|
| **Order ID** | 25728 | CA-2015-SV20365140-42268 |
| **Order Date** | 1430 | 2015-12-31 0:00 |
| **Ship Date** | 1464 | 2016-01-07 0:00 |
| **Ship Mode** | 4 | Standard Class |
| **Customer ID** | 17415 | SV-203651406 |
| **Customer Name** | 796 | Muhammed Yedwab |
| **Segment** | 3 | Consumer |
| **Country** | 165 | United States |
| **City** | 3650 | New York City |
| **State** | 1102 | California |
| **Region** | 23 | Western Europe |
| **Product ID** | 3788 | OFF-FA-6129 |
| **Category** | 3 | Office Supplies |
| **Sub-Category** | 17 | Binders |
| **Product Name** | 3788 | Staples |
| **Market** | 5 | Asia Pacific |
| **Order Priority** | 4 | Medium |

*Table 3 - Orders table numerical variables mean, standard deviation (STD), minimum and maximum values and quantiles (25%, 50%, 75%)*

| | Sales ($) | Quantity | Discount (%) | Profit | Shipping Cost ($) |
|---|---|---|---|---|---|
| **Mean** | 246.491 | 3.477 | 0.143 | 28.611 | 26.479 |
| **STD** | 487.565 | 2.279 | 0.212 | 174.341 | 57.251 |
| **Min** | 0.444 | 1 | 0 | -6599.978 | 1.002 |
| **25%** | 30.759 | 2 | 0 | 0 | 2.61 |
| **50%** | 85.053 | 3 | 0 | 9.24 | 7.79 |
| **75%** | 251.053 | 5 | 0.2 | 36.81 | 24.45 |
| **Max** | 22638.48 | 14 | 0.85 | 8399.976 | 933.57 |

To gain a better understanding of the data in the Orders table, several plots were generated to examine the number of orders according to various factors, including Ship Mode, Segment, Region, Category, and Sub-Category. These plots are presented below


*Figure 1 - Frequency of Orders by Shipping Modes*

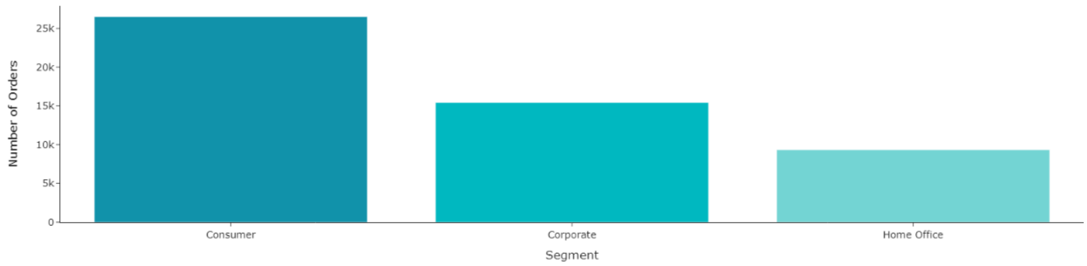


*Figure 2 - Frequency of Orders by Segment*

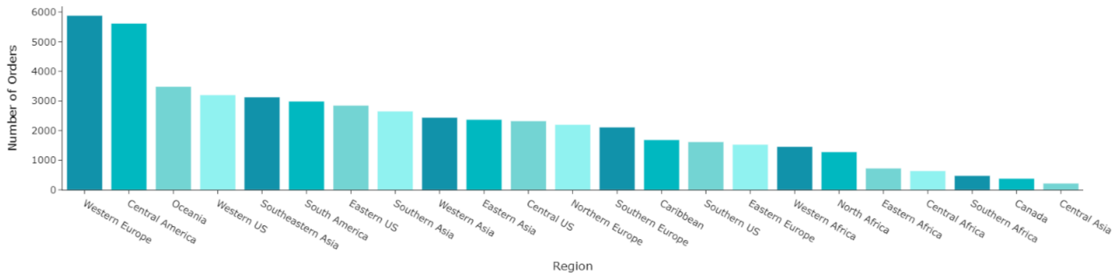


*Figure 3 - Frequency of Orders by Region*

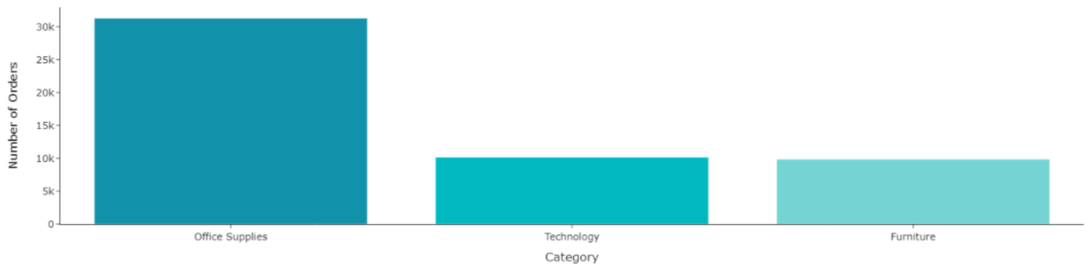


*Figure 4 - Frequency of Orders by Category*

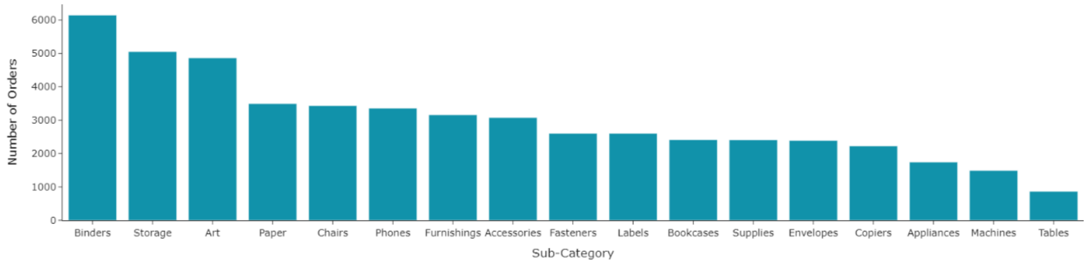


*Figure 5 - Frequency of Orders by Sub-Category*

The above plots offer valuable insights. Firstly, it's evident that Standard Class Shipping is the most preferred mode of shipment amongst the customers. Secondly, the majority of the customers belong to the Consumer Segment. Thirdly, the Central Asia region accounts for the least number of orders, while Western Europe dominates the sales. Moreover, the sales data indicates that Office Supplies is the most popular product category. Lastly, it's interesting to note that Blinders and Paper emerge as clear leaders in sales among customers.

### 1.1.2 - Sellers

The second table of this dataset is the table Sellers. This table contains 24 rows and two columns, which are presented below.

*Table 4 - Sellers table column number (#), name, description and data type*

| # | Column | Description | Data type |
|---|---|---|---|
| **1** | **Person** | Name of the Seller | Object |
| **2** | **Region** | The region where the seller belongs. | Object |

As it is possible to observe in Table 4, the table Sellers contains information regarding the name of the seller and the corresponding location (i.e., region)

To better understand the Sellers table variables, a division was made between categorical and numerical variables so they could be easier described, which can be seen below.

*Table 5 - Sellers table categorical variables description with a column name, number of unique instances, and example of content.*

| Column | Unique | Example |
|---|---|---|
| **Person** | 24 | Marilène Rousseau |
| **Region** | 24 | Caribbean |

Since no numerical variables were present in this table, creating a plot with the provided information would not yield much insight.

## 1.2 - Global Superstore Returns 2016.csv:

This dataset can also be found in the open source "data.world" (https://data.world/tableauhelp/superstore-data-sets). It has a total of 45.03 KB in CSV format, containing one table, Returns. This dataset also contains information regarding the sales of the large-scale e-commerce chain.

### 1.2.1 - Returns

The table Returns displays 1079 rows and three columns, which are presented below.

*Table 6 - Returns table column number (#), name, description and data type.*

| # | Column | Description | Data type |
|---|---|---|---|
| **1** | **Returned** | Boolean that refers if the order was or was not returned | Object |
| **2** | **Order ID** | Unique Order ID for each Customer. | Object |
| **3** | **Region** | The region where the seller belongs | Object |

As it is possible to observe in Table 6, table Returns contains information that shows if a certain order was returned or not, as well as the region of the seller of that order.

For a better understanding of the Returns table variables, a division was made between categorical and numerical variables so they could be easier described, which can be seen below.

*Table 7 – Returns table categorical variables description with a column name, number of unique instances, and example of content.*

| Column | Unique | Example |
|---|---|---|
| **Returned** | 1 | Yes |
| **Order ID** | 1079 | CA-2012-SA20830140-41210 |
| **Region** | 24 | Western Europe |

Since there were no numerical variables present in this table, creating a plot with the provided information would not yield much insight.

## 1.3 - Superstore Dataset

This dataset can be found on the open-source Kaggle website (Superstore Dataset | Kaggle). It has a total of 2.29 MB in CSV format, containing one table, which we will call customers_USA. This dataset also contains information regarding the sales of the large-scale e-commerce chain but now with a focus on the United States of America-based customers.

The table custumers_USA displays 51290 rows and 24 columns, which are presented below

*Table 8 – customers_USA table column number (#), name, description and data type*

| # | Column | Description | Data type |
|---|---|---|---|
| **1** | **Row ID** | Unique ID for each row. | Int64 |
| **2** | **Order ID** | Unique Order ID for each Customer. | Object |
| **3** | **Order Date** | Order Date of the product. | Object |
| **4** | **Ship Date** | Shipping Date of the Product. | Object |
| **5** | **Ship Mode** | Shipping Mode specified by the customer. | Object |
| **6** | **Customer ID** | Unique ID to identify each customer. | Object |
| **7** | **Customer Name** | Name of the Customer. | Object |
| **8** | **Segment** | The segment where the customer belongs. | Object |
| **9** | **Country** | Country of residence of the customer. | Object |
| **10** | **City** | City of residence of the customer. | Object |
| **11** | **State** | State of residence of the customer. | Object |
| **12** | **Postal Code** | Postal Code of every Customer | Int64 |
| **13** | **Region** | The region where the customer belongs. | Object |
| **14** | **Product ID** | Unique ID of the Product. | Object |
| **15** | **Category** | Category of the product ordered. | Object |
| **16** | **Sub-Category** | The sub-Category of the product ordered. | Object |
| **17** | **Product Name** | Name of the Product | Object |
| **18** | **Sales** | Sales of the Product. | Float64 |
| **19** | **Quantity** | Quantity of the Product. | Int64 |
| **20** | **Discount** | Discount provided. | Float64 |
| **21** | **Profit** | Profit/Loss incurred. | Float64 |

As it is possible to observe in Table 8, the table customers_USA contains information similar to the Orders table, which goes from details about the dates of the orders and shipments, products categories and profits, with the difference that this table doesn't contain information regarding the seller, being its focus on the customer (USA based), providing more information, such as region, city, state, name, etc.

To better understand the customers_USA table variables, a division was made between categorical and numerical variables so they could be easier described, which can be seen below.

*Table 9 – customers_USA table categorical variables description with a column name, number of unique instances, and example of content.*

| Column | Unique | Example |
|---|---|---|
| **Order ID** | 5009 | CA-2017-100111 |
| **Order Date** | 1237 | 9/5/2016 |
| **Ship Date** | 1334 | 12/16/2015 |
| **Ship Mode** | 4 | Standard Class |
| **Customer ID** | 793 | WB-21850 |
| **Customer Name** | 793 | William Brown |
| **Segment** | 3 | Consumer |
| **Country** | 1 | United States |
| **City** | 531 | New York City |
| **State** | 49 | California |
| **Region** | 4 | West |
| **Product ID** | 1862 | OFF-PA-10001970 |
| **Category** | 3 | Office Supplies |
| **Sub-category** | 17 | Binders |
| **Product Name** | 1850 | Staple envelope |
| **Postal Code** | 631 | 42420 |

*Table 10 - customers_USA table numerical variables mean, standard deviation (STD), minimum and maximum values and quantiles (25%, 50%, 75%)*

| | Sales ($) | Quantity | Discount (%) | Profit ($) |
|---|---|---|---|---|
| **Mean** | 229.858 | 3.789 | 0.156 | 28.657 |
| **STD** | 623.245 | 2.225 | 0.206 | 234.26 |
| **Min** | 0.444 | 1 | 0 | -6599.978 |
| **25%** | 17.28 | 2 | 0 | 1.729 |
| **50%** | 54.49 | 3 | 0.2 | 8.667 |
| **75%** | 209.94 | 5 | 0.2 | 29.364 |
| **Max** | 22638.48 | 14 | 0.8 | 8399.976 |

Some plots were made to better understand the shape of the data present in the customers_USA table. The number of customers was analysed by State, City and Region. The plots can be seen below.


*Figure 6 - Number of Customers by State Map*

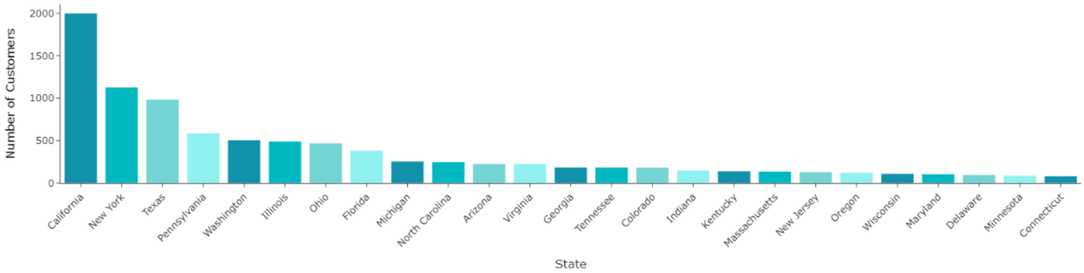


*Figure 7 - Number of Customers by State (Top 25)*

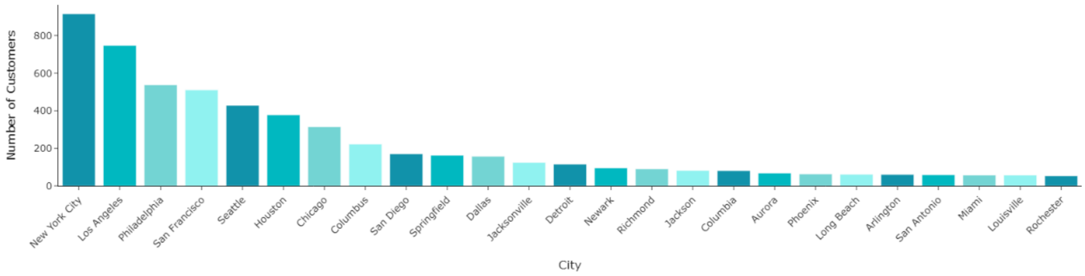


*Figure 8 - Number of Customers by City (Top 25)*

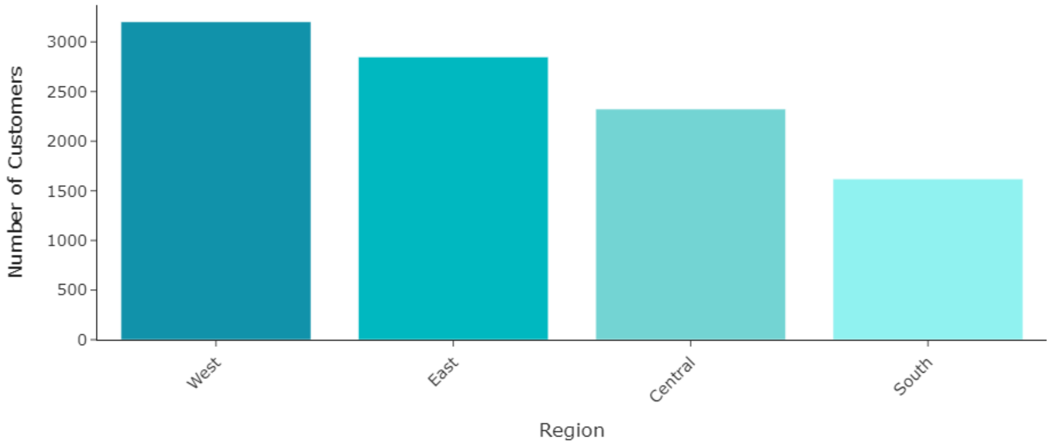


*Figure 9 - Number of Customers by Region*

Regarding these plots, three orders of insights can be taken. The three most popular states among the customers are California, New York and Texas, and the least is Connecticut. The most popular city among the customers is New York City, and the 25th is Rochester. And most popular region among customers is the West region, and the least popular is the South region.

## 1.4 - US Holiday Dates (2004 - 2021)

This dataset can be found on the open-source Kaggle website (US Holiday Dates (2004 - 2021) | Kaggle). It has a total of 15.7 kB in CSV format, containing one table, which we will call holiday_USA, with information regarding the dates of US holidays.

The table holiday_USA displays 342 rows and six columns, presented below.

*Figure 10 holiday_USA table column number (#), name, description and data type*

| # | Column | Description | Data type |
|---|---|---|---|
| **1** | **Date** | Date corresponding to holiday | Object |
| **2** | **Holiday** | Holiday designation | Object |
| **3** | **WeekDay** | Day of the week when the holiday occurs | Object |
| **4** | **Month** | The month when the holiday occurs | Int64 |
| **5** | **Day** | Day of the month when the holiday occurs | Int64 |
| **6** | **Year** | Year of reference [2004-2021] | Int64 |

As it is possible to observe in Table 10, table holiday_USA contains a list of holidays that includes 18 years of US Holidays dated between 2004 and 2021. Each record has a Date, Holiday, Weekday, Month, Day and Year.

For a better understanding of the holiday_USA table variables, a division was made between categorical and numerical variables so they could be easier described, which can be seen below.

*Table 11 - holiday_USA table categorical variables description with a column name, number of unique instances and example of content.*

| Column | Unique | Example |
|---|---|---|
| **Date** | 336 | 2007-04-08 |
| **Holiday** | 18 | Labor Day Weekend |
| **WeekDay** | 7 | Monday |

Since there were no numerical variables present in this table, creating a plot with the provided information would not yield much insight.

## 1.5 - Highest GDP Counties in the USA

This dataset can also be found on the open-source Kaggle website (Highest GDP Counties in USA | Kaggle). It has a total of 3.65 MB in CSV format, containing one table, which we will call GDP_USA. This dataset also includes data on the GDP of counties in the United States

The table GDP_USA displays 55501 rows and eight columns, presented below.

*Table 12 - GDP_USA table column number (#), name, description and data type.*

| # | Column | Description | Data Type |
|---|---|---|---|
| **1** | **index** | | Int64 |
| **2** | **Year** | Year of reference | Int64 |
| **3** | **Region** | The region of reference for the GDP | Object |
| **4** | **SUB_REGION** | The sub-region of reference for the GDP | Object |
| **5** | **State** | The state of reference for the GDP | Object |
| **6** | **County** | The county of reference for the GDP | Object |
| **7** | **GDP(Chained $)** | GDP value in $ | Float64 |

As it is possible to observe in Table 12, the table GDP_USA contains data on the GDP in dollars, of counties in the United States, organised by year, region, sub-region, and county

To better understand the GDP_USA table variables, a division was made between categorical and numerical variables so they could be easier described, which can be seen below.

*Table 13 - GDP_USA table categorical variables description with a column name, number of unique instances and example of content.*

| Column | Unique | Example |
|---|---|---|
| **Year** | 18 | 2001 |
| **Region** | 8 | Southeast |
| **SUB_REGION** | 9 | West North Central |
| **State** | 51 | Texas |
| **County** | 1801 | Washington |

*Table 14 - GDP_USA table numerical variables mean, standard deviation (STD), minimum and maximum values and quantiles(25%, 50%, 75%).*

| | GDP (Chained $) |
|---|---|
| **Mean** | 5.05E+09 |
| **STD** | 2.15E+10 |
| **Min** | 9.95E+06 |
| **25%** | 3.37E+08 |
| **50%** | 8.61E+08 |
| **75%** | 2.49E+09 |
| **Max** | 7.11E+11 |

To better understand the shape of the data present in the GDP_USA table, some plots were done. The GDP (Chained $) was analysed by State and Region. The plots can be seen below.

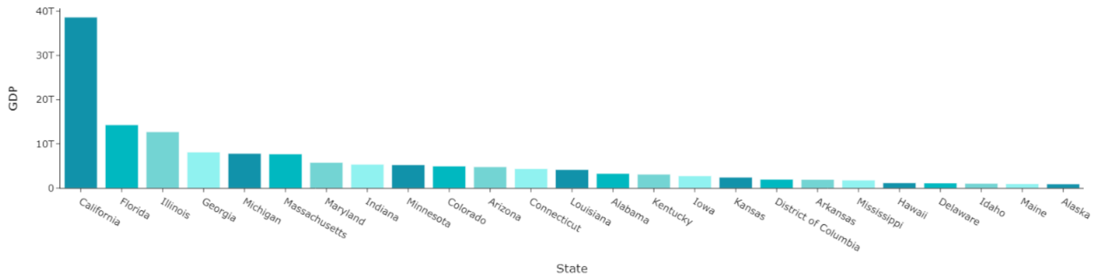


*Figure 11 GDP by State (Top 25)*

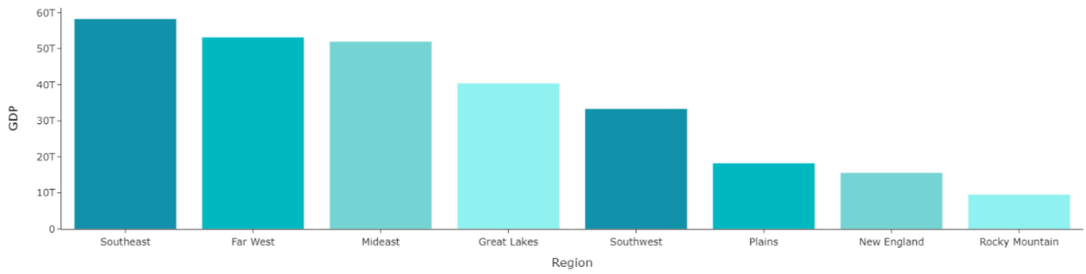


*Figure 12 GDP by Region*

It is possible to observe that the state with the highest GDP is the state of California, and the 25th is Alaska. In terms of regions, the one with the highest GDP is the Southeast, and the one with the lowest is the Rocky Mountains.sss

# 2. Analyse values and errors in the data fields of each source

Data standardisation is a critical step in data analysis and pre-processing (Gal & Rubinfeld, 2018). It involves bringing different data sources into a common format, allowing for meaningful comparisons and accurate analysis (Aggarwal, 2015). A critical aspect of data standardisation is analysing each feature's values and errors of all the used data sources. This involves carefully examining the data to identify any inconsistencies or anomalies and ensuring that the data is reliable and accurate (Gal & Rubinfeld, 2018). By standardising the data, organisations can make better-informed decisions based on consistent, reliable data, leading to improved efficiency and better outcomes (Gal & Rubinfeld, 2018).

Performing the analysis of each feature's values and errors of all the used data sources is a crucial step in data standardisation, and this can be achieved using programming languages such as Python. Python is a widely used language for data analysis, with numerous libraries and tools available that make it an ideal choice for standardising data.

In this context, the analysis of each feature's values and errors will cover all four tables, namely the orders, returns, sellers, customers_USA, holiday_USA and GDP_USA tables. These tables contain different types of data, including transactional data, customer data, and sales data. The analysis will involve checking for any missing or inconsistent data, identifying any outliers or anomalies, and addressing any issues that may affect the data quality.

Using Python for this analysis makes it possible to automate the process and perform it more efficiently. This ensures that the standardisation of data is not only accurate but also timely. Additionally, using Python makes it easier to visualise and communicate the results of the analysis, allowing for better decision-making based on reliable data.

In some cases, a reoccurring situation that occurs in the tables is that some features are mistakenly represented as object data types when they should be represented as a different kind of data type. Object data types can be used for a variety of data, including text and alphanumeric characters. However, in certain scenarios, they may not be the most efficient or appropriate data type to use.

Using the wrong data type can result in a range of issues, including increased memory usage, reduced processing speed, and decreased accuracy (Tabatabaei et al., 2021). For example, if a feature such as a date is represented as an object data type instead of a datetime data type, it may be challenging to perform certain types of analysis on the data, such as sorting or filtering by date.

## 2.1 - Orders table

With the goal of standardising the data, a thorough analysis was performed on the table orders, which revealed several data type inconsistencies. First, some features that were meant to be dates were represented as object data types, prompting a change to the appropriate date format. Similarly, all information regarding the IDs (Order ID and Customer ID) was classified as an object data type, but to avoid potential issues stemming from mixed numerical and alphabetical characters, it was converted to a string data type. The same process was applied to other features like the City, State, Country, Region, Market, Product Name

Furthermore, there are other columns in the 'orders' table, such as Segment, Category, Sub-Category, and Order Priority, that are currently represented as an object data type but could benefit from a change to a category data type. This type of conversion can offer performance and memory benefits, as well as make it easier to perform certain types of analysis on the data.

Lastly, it was found that the postal code column had more than 50% of missing data. Since this column might not be used for the analysis, it was discarded from the dataset. The reordering columns had a proper data type classification.

### 2.2 - Customers table

Since the customer's table has the same kind of data type inconsistencies as the orders table, the same modifications were performed to standardise the data.

### 2.3 - Returns table

To improve performance and consistency, certain modifications were made to the returns table. Specifically, the "Returned" column, which previously consisted of "Yes" or "No" values, was replaced with a Boolean data type. Similarly, the Order ID in the returns table was changed to a string data type, following the same reasoning applied to the orders table. As for the Region column, instead of it being an object data type it was changed to a string data type.

### 2.4 - Sellers table

In the sellers table, the Person and Region columns were initially of the object data type. They were transformed to string data type to ensure consistency with the rest of the data.

### 2.5 - Holiday USA table

Like the previous tables, it was essential to modify the data types of certain columns in this table to ensure consistency with the rest of the data. The Date column was converted to a date data type to facilitate date-based computations. Additionally, the Holiday and WeekDay columns, which were initially of object data type, were converted to string type to ensure consistency across all columns.

To work with a subset of the data that is consistent with the orders dataset, only the rows corresponding to the years 2012 to 2015 were selected. Additionally, the Date column was in a different date format than the one used in the orders dataset. Therefore, the date format was converted from Y-M-D to D-M-Y to ensure consistency across all columns.

### 2.6 - USA GDP table

Several modifications were made to the GDP table to ensure consistency with the orders dataset and enable better data analysis. Firstly, the unnecessary columns, namely Region, SUB_REGION, and County, were removed as they do not provide any useful information for the analysis.

Furthermore, it was observed that the orders dataset did not contain data for the states of Alaska and Hawaii, while the GDP table did. To orderstain consistency between the datasets, the data for Alaska and Hawaii were removed from the GDP table.

To gain a better understanding of the data associated with counties, the data was grouped by state. Two new columns were created to provide insights into the GDP of each state: Avg_GDP column containing the average GDP of each state, and Total_GDP column containing the sum of the GDP of all counties within each state.

Finally, the State column was converted to a string data type to ensure consistency with the rest of the data.

## 3 - Draw a diagram with links between data sources

Once all the necessary information was collected, the next step was to create a diagram that would effectively illustrate the relationships between the data. Figure 17 presents the resulting diagram, which displays all the relevant tables, their corresponding variables, and the connections between the tables that are established through these variables.

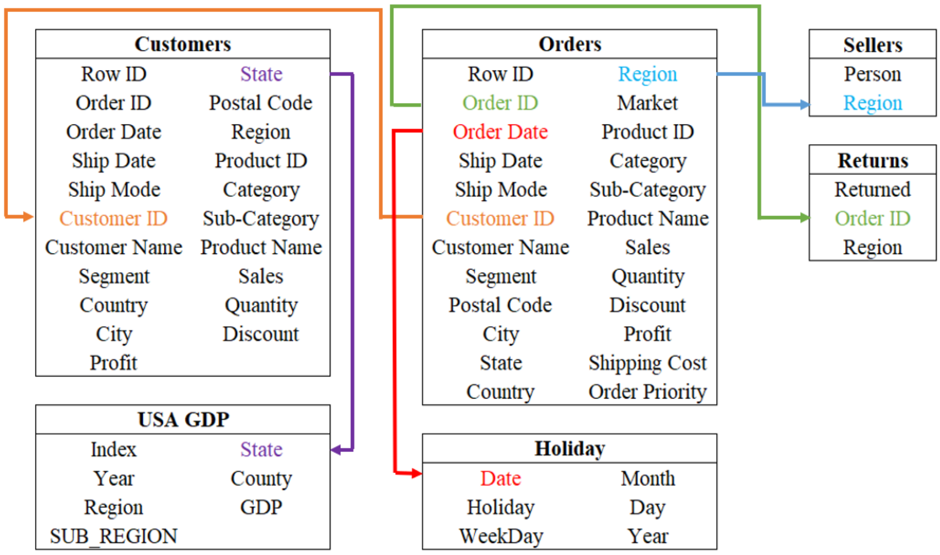


*Figure 13 - Diagram with all tables chosen for this project as well as their connections.*

# 4 - Describe a business process to take advantage of data

The datasets presented simulate a typical sales process that is common to many businesses. They contain valuable information on product sales, customer and store demographics, orders and shipping, product returns, and promotional activity. By analysing this data, businesses can gain insights into their performance over time and across different locations, and identify areas for improvement. As an example, an analyst could use the dataset to:

- Analyse sales trends over time and identify top-selling and most profitable products or categories. These values can also be examined across different locations to identify underperforming areas and determine which items to promote in each.
- Identify the customer segments that generate the most revenue, their geographical locations, preferred products, and how promotions influence their purchasing behavior. This information can be used to adjust marketing strategies and optimise promotions to increase revenue and customer satisfaction.
- Evaluate shipping performance across different stores and identify areas for improvement in delivery times and shipping costs. Analyse if these factors are linked to low sales or high refund counts for other products.
- Investigate the root causes of poor delivery times and high shipping costs, such as the distance from the store to the customer, the type of product, order priority, or delivery mode. This information can be used to improve delivery times and lower shipping costs, thereby enhancing customer trust and satisfaction.
- Utilising the orders information, we can identify which products are often purchased together and determine how marketing strategies on certain products affect the sales of others. This information can be used to optimise inventory and strategise which promotions would be most effective, while also identifying products that are not influenced by promotions and adjusting marketing accordingly. Additionally, we can determine which products have seasonal sales patterns and target them during the most profitable timeframes.
- We can also analyse trends in customer loyalty across different locations by examining repeat purchases and identifying what top-performing sellers do differently. This information can be used to improve customer trust and satisfaction and inform strategies for improving sales in underperforming locations.

Given the project's scope, this analysis will focus on one critical business process: product profitability.

Overall, these datasets can be used to gain insights into the sales process of a business and identify opportunities for optimisation and growth.

# 5 - Define three analytical questions for the business process

According to these datasets, it would be beneficial to understand which products, regions, categories, and customer segments this chain should target or avoid as well as the best and worst performing sellers. For that, three analytical questions aligned with the orders of three components that make up our business process (product profit) were created.

## 5.1 - Sellers Analysis

There are various questions relative to the seller that can be asked which are aligned with our business process (product profit).

Which seller generates the highest / lowest profit? This can be considered the orders question, from which we can start decomposing. Who is the seller with the highest or lowest profit? Other variables can be analysed to help answer this question, for example: Do the customers of the best seller belong to a state/city/region with high GDP values? The inverse can be asked for the worst seller as well. What is the most frequent customer segment and product category of the highest and lowest seller? Do discounts influence the seller's profit? All these questions can give various insights into the influences or at least what correlates with the highest and lowest sellers.

## 5.2 - Customer analysis

Which customers are most valuable and profitable? Which ones keep returning to the same store(seller)? Do customers increase confidence in the stores with time? For example, do customers increase the frequency and or amount spent on a particular store(seller)?

What is the percentage of customers, out of all customers of the superstore, that returned a product? What was the shipment mode that customers preferred the least for every region the superstore operates in? Are most of the orders of products of the "Furniture" category made by customers with the consumer or corporate segment? Does the customer importance relate to how often they ask for critical priority orders? And being more specific: How many customers that ordered products from the Europe market conducted critical priority orders in the year 2015?

These questions offer decision-makers greater insight into customer behaviour which can be critical to improving the business process of the superstore. Information such as frequency and distribution of the orders and returns, the number of individuals who were always satisfied with their purchases and the individuals who had issues with what they ordered on multiple occasions. Putting these instances under a microscope and studying the reason for them can be very benefitting for the superstore.

## 5.3 - Product analysis

What was the season of the year when orders for "GlobeWeis Clasp Envelopes" were made in higher amounts? Does this or any other product orders relate to the occurrence of holidays? What were the top 3 categories and respective sub-categories of products that generated the most profit for the superstore? For these three categories, what was the sub-category that had products with the highest discounts? Was there any crossover between the sub-categories for this answer and the sub-categories that generated the highest profit asked before? What is the average price of shipping that made customers not buy the product?

These questions, more centred around the products being sold, allow the decision makers to see clearly at what moments in a year a product is in higher demand and if they would benefit more from increased discount percentages at those times. What categories and sub-categories are the most profitable for the company is also indispensable information for the people tasked with deciding prices and quantities of future resupply orders. To be able to single out less profitable products is of worth to the superstore as studies can be made as to why these products are not performing as well as the rest and if needed discontinuation from selling them all together can now be supported from this analysis made available from being able to answer questions such as the ones asked here.

# 6 - Bibliography

Aggarwal, C. C. (2015). *Data Mining*. Springer International Publishing. https://doi.org/10.1007/978-3-319-14142-8

Gal, M., & Rubinfeld, D. (2018). Data Standardisation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3326377

Tabatabaei, S.-K., Fatahi Valilai, O., Abedian, A., & Khalilzadeh, M. (2021). A novel framework for storage assignment optimisation inspired by finite element method. *PeerJ Computer Science*, *7*, e378. https://doi.org/10.7717/peerj-cs.378